*Current Topics*

## *In Silico* Technologies to Boost Pharmaceutical Development

*Regular Article*

# A Data-Driven Approach to Predicting Tablet Properties after Accelerated Test Using Raw Material Property Database and Machine Learning

Yoshihiro Hayashi,*[a,b] Yuri Nakano,[b] Yuki Marumo,[b] Shungo Kumada,[a] Kotaro Okada,[b] and Yoshinori Onuki[b]

[a] *Pharmaceutical Technology Division, Nichi-Iko Pharmaceutical Co., Ltd., 205–1, Shimoumezawa, Namerikawa, Toyama 936–0857, Japan: and* [b] *Department of Pharmaceutical Technology, Graduate School of Medicine and Pharmaceutical Science for Research, University of Toyama, 2630 Sugitani, Toyama 930–0194, Japan.*

Received July 25, 2022; accepted October 3, 2022

**The purpose of this study was to develop a model for predicting tablet properties after an accelerated test and to determine whether molecular descriptors affect tablet properties. Tablets were prepared using 81 types of active pharmaceutical ingredients, with the same formulation and three different levels of compression pressure. The tablet properties measured were the tensile strength and disintegration time of tablets after two weeks of accelerated test. The material properties measured were the change in tablet thickness before and after the accelerated test, maximum swelling force, swelling time, and swelling rate. The acquired data were added to our previously constructed database containing a total of 20 material properties and 3381 molecular descriptors. The feature importance values of molecular descriptors, material properties and the compression pressure for each tablet property were calculated by random forest, which is one type of machine learning (ML) that uses ensemble learning and decision trees. The results showed that more than half of the top 25 most important features were molecular descriptors for both tablet properties, indicating that molecular descriptors are strongly related to tablet properties. A prediction model of tablet properties was constructed by eight ML types using 25 of the most important features. The results showed that the boosted neural network exhibited the best prediction accuracy and was able to predict tablet properties with high accuracy. A data-driven approach is useful for discovering intricate relationships hidden within complex and large data sets and predicting tablet properties after an accelerated test.**

**Key words** machine learning, material library, data-driven, molecular descriptor, quantitative structure–property relationship, tablet

## Introduction

In the design of oral solid dosage forms, it is important to optimize various design variables to consistently produce dosage forms that meet multiple quality attribute specifications. For example, tablet design variables include formulations such as the type and ratio of excipients, and the manufacturing method such as the type of manufacturing equipment, process flow, and process parameters.[1)] Because there are more than 1000 types of excipients, there are a vast number of possible combinations of excipients. The relationship between these design variables and quality characteristics is a complex one, in which various factors interact. Quality characteristics often also have a trade-off relationship with each other, and thus it is extremely difficult to design a product so that multiple quality attributes satisfy the standard. In addition, the relationship between these design variables and quality attributes varies significantly depending on the type and amount of the active pharmaceutical ingredients (APIs) to be compounded,[2)] so that pharmaceutical design must be performed each time a new drug candidate compound is created. Against this background, predicting the optimal design for the solid dosage form is difficult and still relies heavily on the traditional trial-and-error approach of pharmaceutical researchers.[3)]

One way for breaking out of the traditional trial-and-error approach is the data-driven approach. This approach is to make decisions, solve problems, and gain new knowledge based on big data, which is a vast store of information of various types, and the results of analysis processed by machine learning (ML) algorithms. Data-driven approaches have already been studied in many other fields, including materials science.[4,5)] In the traditional approach, new materials are discovered by experiment, theory, and computation. In data-driven materials science, by contrast, databases are built and new materials are discovered through ML. For example, if one wants to predict the physicochemical properties of a particular substance, a database is constructed by acquiring various characteristics, so-called features, about the substance. The features may include molecular descriptors and, in the design of oral solid dosage forms, material properties, formulations, manufacturing methods, and information obtained from process analysis techniques. By modeling the acquired data through ML, it is possible to predict a specific physicochemi-

* To whom correspondence should be addressed. e-mail: yoshihiro-hayashi@nichiiko.co.jp

cal property from the features and to identify the importance of the features. Furthermore, it predicts the composition and manufacturing process that will enable the preparation of a product with the desired properties.

In the field of the design of oral solid dosage forms, there have been several reports on data-driven approaches. In this field, based on the concept of quality by design, small datasets are constructed mainly according to the design of experiments, and multiple regression analysis or partial least squares (PLS) has been used to model causal relationships.[6–8] In recent years, several cases have been reported in which databases containing more information than ever before have been constructed and complex relationships have been modeled using more powerful ML.[9–16] For instance, Paul *et al.* produced 26 different types of tablets with different excipient types and amounts and different tableting pressures.[17] They showed by ML such as PLS and random forest (RF) that the compaction parameters and material properties were strongly related to capping tendency. Galata *et al.* prepared 56 different tablets with different percentages of API, and hydroxypropyl methylcellulose (HPMC), compression force, and HPMC particle size fractions.[18] They showed that ML types such as artificial neural networks (ANN), support vector machines (SVM), and ensemble of regression trees can accurately predict dissolution profiles from near-IR spectra, tableting pressure, and HPMC particle size distribution. Takayama *et al.* prepared 112 types of tablets with different types and API amounts.[19] They showed that a four-layered ANN can accurately predict the dissolution profiles based on the physicochemical properties of the API and the powder properties.

In earlier work, we constructed a material library containing 3381 types of molecular descriptors, 20 types of material properties, and 3 types of tablet properties (tensile strength (TS), disintegration time (DT), and tablet density) for 81 APIs,[20–22] being one of the largest raw material property databases in terms of API types and tablet properties. Then, we applied RF and boosted tree (BT) to our material library and developed a model to predict the TS and DT values of tablets from the API characteristics and compression pressure. We have shown that RF and BT could predict tablet properties more accurately than PLS, which is commonly used in pharmaceutical development, and that molecular descriptors are closely related to the tablet density and the true density of the API. Based on these studies, in the current work we focus on the following three issues.

(1) Expansion of the material library. The storage stability of tablets is one of the fundamental properties in pharmaceutical development. If the tablet properties after storage can be predicted, the development period could be reduced significantly. However, to our knowledge, there is no database that contains post-storage tablet characterization data for a wide variety of APIs. Takagaki *et al.* have constructed a database regarding the physicochemical properties of APIs, tablet hardness, and DT values after the accelerated test, but only five API types are available.[23] Therefore, we newly prepared 243 types of tablets with different compression pressures for 81 API types, and measured the TS and DT after storage. In addition, we measured the change in tablet thickness before and after the accelerated test and the swelling characteristics, which is one of the properties strongly related to the disintegration mechanism,[24] and the obtained data was added to our previously constructed material library.

(2) Comparison of ML types. We have applied two ML types, namely RF and BT, which use ensemble learning and decision trees.[21] Because the best algorithm is case-by-case and is highly dependent on multiple factors relevant to datasets and objectives,[11] we also focused on the following ML types: boosted neural network (BNN), SVM, k-nearest neighbor algorithm (kNN), Ridge regression, least absolute shrinkage and selection operator (LASSO), and elastic net. These ML techniques have been widely applied in other fields,[25–29] and may have higher prediction accuracy than BT and RF. However, they have rarely been applied in the field of oral solid dosage form development.[11]

(3) Effect of molecular descriptors on tablet properties. Molecular descriptors are molecular properties computationally calculated from molecular structures and are characteristic values related to molecular features.[30] They have been used in fields such as quantitative structure–property relationships and have been shown to be useful in predicting various physical properties.[31] We have also reported their usefulness in predicting tablet density and the true density of the API,[22] and they may be relevant to other tablet properties as well. However, there are still no papers evaluating the effect of molecular descriptors on tablet properties. A database containing a wide variety of APIs is needed to evaluate the influence of molecular descriptors, but such a database does not exist in the field of pharmaceutical science, and thus has not been examined.

The purpose of this study was to develop a model of predicting tablet properties after the accelerated test and to determine whether molecular descriptors affect tablet properties. In this study, new material properties were added to our previously constructed database. To evaluate the relationship between molecular descriptors and tablet properties, feature importance was calculated using RF. In addition, models with and without molecular descriptors were constructed using multiple ML methods, and the prediction accuracies were compared. This allowed us to evaluate how the prediction accuracy of tablet properties changes depending on the molecular descriptors.

## Experimental

**Materials** Eighty-one types of model APIs were purchased from FUJIFILM Wako Pure Chemical Corporation (Osaka, Japan), Tokyo Chemical Industry Co., Ltd. (Tokyo, Japan), or Yamamoto Corporation Co., Ltd. (Osaka, Japan). Several APIs were ground in a mortar and pestle to reduce their particle size (as they were too large for direct compression). Microcrystalline cellulose (MCC; Ceolus PH-101, Asahi Kasei Chemicals Co. Ltd., Tokyo, Japan) and magnesium stearate (Mg-St; FUJIFILM Wako Pure Chemical Corporation) were purchased from commercial suppliers.

**Preparation of Tablets** The tablets contained 100 mg API, 98 mg MCC, and 2 mg Mg-St (50% API, 49% MCC, and 1% Mg-St). A mixture of MCC and Mg-St was prepared for 81 APIs, and then the mixture was mixed with each API, *i.e.*, 343 g of MCC and 7 g of Mg-St were mixed in a polyethylene (PE) bag for 1 min, and 2.5 g of the mixture was mixed with 2.5 g of API in a PE bag for 1 min. The final blend was directly compressed into round-faced tablets (200 mg, 8 mm diameter and 12 mm radius of curvature) using a tableting machine equipped with a load cell (AUTOTAB-100W, Ichihashi-

Seiki Co., Ltd., Kyoto, Japan). Samples were manually filled into the die, and the compression pressure was set at 120, 160, and 200 MPa.

**Evaluation of TS, DT, and Change in Tablet Thickness after Accelerated Test** The accelerated test was performed by storing the tablets at 40 °C and 75% relative humidity for two weeks in a stability chamber (CSH-110; ESPEC Company, Osaka, Japan). The samples were then stored in a desiccator at room temperature for at least 24 h. After that, the TS and DT were measured by the following methods.

The hardness and thickness of the tablets were determined with a tablet hardness tester (Portable checker PC-30; Okada Seiko, Tokyo, Japan) and a digital micrometer caliper (MDQ-30M, Mitutoyo Corporation, Tokyo, Japan), respectively. The TS was determined according to the following equation[32]:

$$\mathrm{TS} = \frac{10F}{\pi D^2}\left[\frac{2.84t}{D} - \frac{0.126t}{W} + \frac{3.15W}{D} + 0.01\right]^{-1} \quad (1)$$

where $F$ is the maximal diametrical crushing force, and $D$, $t$, and $W$ are the tablet diameter, tablet thickness, and the thickness of the band part, respectively. The TS value of each API was measured in triplicate. Each set of data reported is the average of the triplicate measurements.

The change in tablet thickness before and after the accelerated test were calculated according to the following equation:

$$\textit{Change in tablet thickness}\ (\%) = 100 \times \frac{(t_{after} - t_{before})}{t_{before}} \quad (2)$$

where $t_{before}$ and $t_{after}$ are thickness before and after the accelerated test, respectively.

The disintegration test was performed according to the Japanese Pharmacopoeia, 18th edition (JP18) disintegration test for tablets, using a disintegration tester (NT-20H; Toyama Sangyo Co., Ltd., Osaka, Japan) and purified water (as solvent) at 37 ± 2 °C. Discs were not used. DT was defined as the interval required for the complete disappearance of a tablet or its particles from the tester net. The DT values of each of the APIs were measured in triplicate. Each set of data reported is the average of triplicate measurements. The logarithmic transformation of DT ($\log DT$) was applied for normalization. The $\log DT$ values of the tablets that did not disintegrate within 30 min were taken as 7.495 (DT = 30 min) in the following analysis.

**Evaluation of Swelling Behavior** The swelling behavior of the tablets was measured using NEW GRANO (Okada Seiko) according to a published method,[33] with some modification. The sponge, mesh, tablets, and lid were set inside a cylindrical tube in that order. The parts were then immersed in a water bath, and as the tablets began to swell, the load generated by this on the lid was measured over time with a load cell. For the swelling behavior measurements, flat-faced tablets compressed at 200 MPa with a tableting machine (AUTOTAB-100 W, Ichihachi-Seiki Co., Ltd.) were used. Swelling behavior varies with tableting pressure, but it is difficult to evaluate the relationship between these two parameters (tableting pressure and swelling behavior) because of the enormous amount of time required for measurement. Therefore, swelling behavior was evaluated only for tablets with a compression pressure of 200 MPa. The formulation was the same as that used for the hardness and disintegration time measurements. Swelling behavior was evaluated for tablets before the accelerated test. The maximum swelling force, swelling time, and swelling rate were calculated from the measurements. The swelling time is the time required to generate a load of half the maximum swelling force, and the swelling rate is the value obtained by dividing the maximum swelling force by twice the swelling time, referring to previous studies.[33] The swelling behavior was measured until equilibrium was reached, but the maximum measurement time was limited to 30 min because the swelling force continued to increase slightly for some APIs. Because the time of swelling onset differed significantly according to API, the starting point was defined as the time when a load of approximately 1 N was applied. Eighty-one types of our database have been constructed so far, but only 73 types of tablets were measured for swelling behavior in this study due to limited raw materials and time. When building the ML model, to compensate for missing values for the APIs that could not be measured because they did not follow a normal distribution, we used the median of all data. Swelling behavior measurements were taken three times, and the average value of each swelling property was calculated. For two types of APIs (streptomycin sulfate and norfloxacin), it was not possible to perform the experiment three times and the number of measurements was limited to two.

**Constructed Database** The newly measured data from this study were added to our previously constructed database.[20–22] The database contains information on 81 API types. A summary is given below.

(1) The 24 types of material properties included: change in tablet thickness, maximum swelling force, swelling rate, swelling time, API form (salt or free form), powder particle diameters at the 10th, 50th, and 90th percentiles of the cumulative percent undersize distribution ($d10$, $d50$, and $d90$), modal diameter, span ($(d90-d10)/d50$), bulk density and tapped density, Hausner ratio, loss on drying, in-die elastic recovery, true density, solubility, hygroscopicity, water adsorption rate, total surface energy, polar surface energy, and dispersive surface energy. Molecular weight and the partition coefficient ($\log P$) were obtained from PubChem (https://www.ncbi.nlm.nih.gov/pccompound). Although molecular weight, $\log P$, and solubility are close to being molecular descriptors, they were considered material properties in this study.

(2) Among the 3381 types of molecular descriptors were two-dimensional (2D) molecular descriptors generated using Dragon (v7.0.8, Talete srl, Milano, Italy). The descriptors included various constitutional indices, ring descriptors, topological indices, connectivity indices, extended topochemical atom indices, and others.

(3) Process parameters: the compression pressure at three levels, namely 120, 160, and 200 MPa.

(4) The tablet properties included were TS and DT before and after the accelerated test.

**Construction of Prediction Model for TS and DT after Accelerated Test Using ML** Eight ML types, *i.e.*, BT, RF, BNN, SVM, kNN, Ridge regression, LASSO, and elastic net, were performed using the Model Screening platform in JMP Pro (version 16, SAS Institute Inc., Cary, NC, U.S.A.). Excellent literature has been published on ML,[11,29] and therefore we have omitted details in this paper.

TS or $\log DT$ after the accelerated test was selected as an objective variable in ML. A large number of explanatory variables can lead to enormous computation time and

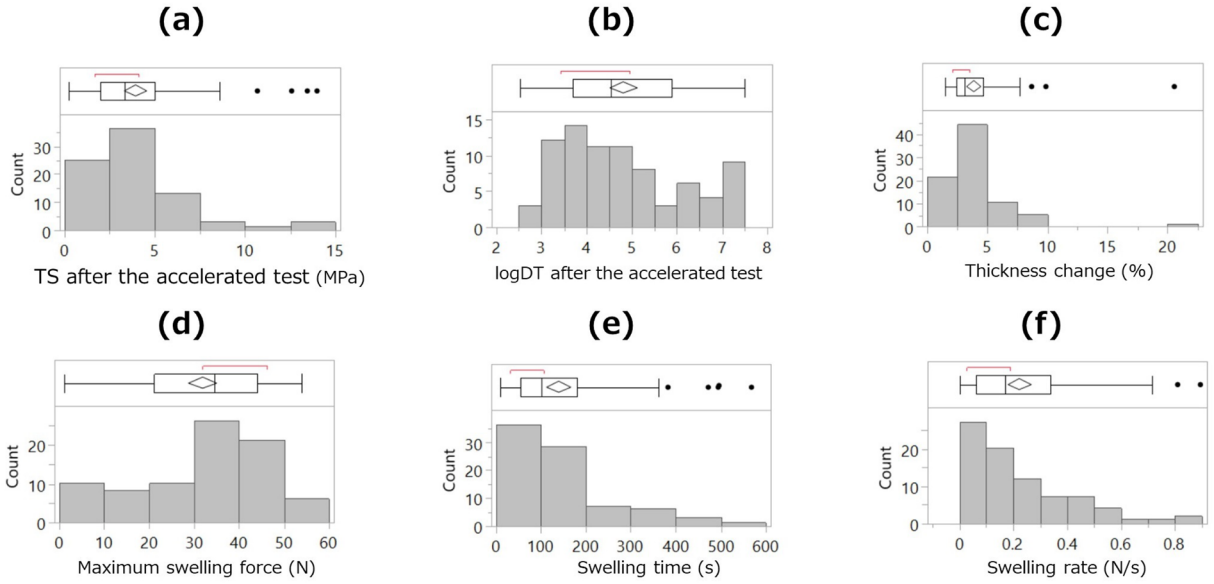


Fig. 1. Histograms and Box Plots of (a) TS and (b) Log*DT* after the Accelerated Test, (c) Thickness Change, (d) Maximum Swelling Force, (e) Swelling Time, and (f) Swelling Rate

The center of the diamond in the box plot represents the mean, and the two ends of the diamond represent the 95% confidence interval of the mean. The square brackets shown outside the box plot indicate the shortest range where 50% of the data is dense.

overlearning; therefore, only the 25 features with the highest importance were used as explanatory variables (see next section for feature importance). In linear models such as the Ridge regression, LASSO, and elastic net, a model was also constructed in which the squared and interaction terms of the features were included as explanatory variables. All explanatory variables were standardized to make the scales uniform.

**Estimation of Feature Importance** The feature importance values for TS and log*DT* after the accelerated test were calculated based on RF, using the predictor screening platform in JMP Pro (version 16, SAS Institute Inc.). RF was selected as the feature selection method because of its relatively low computational cost among nonlinear ML methods and its clear method of calculating feature importance. That is, the effect of a total of 3406 features including 24 material properties, 3381 molecular descriptors, and the compression pressure on each tablet property was calculated. The number of decision trees in the model was set to 1000.

**Prediction Accuracy of the Constructed Model** Each model was constructed using holdout validation. Data sets were split into the following three subsets: training, validation, and test sets. The training set consisted of 60% of all cases and was used for model construction. The validation set (20% of all cases) was required to determine the optimal hyperparameters. The test set (20% of all cases) was required to evaluate the predictive ability of the unknown samples. Using stratified random sampling, the sample was randomly divided into a training set, a validation set, and a test set. That is, the data were divided so that the distribution of the objective variables (TS and log*DT* after the accelerated test) was identical in all subsets. The stratified random sampling method was repeated five times, and five kinds of subsets were created to evaluate the extent of variation in the prediction accuracy of the model and the relative importance of the explanatory variables according to the combination of the selected samples. $R^2$ and the root mean square error (RMSE) were calculated to quantitatively evaluate prediction accuracy.

## Results and Discussion

**Distribution of Properties Newly Added to the Database** To characterize the six new properties added to our database, *e.g.*, TS and DT after the accelerated test, maximum swelling force, swelling time, swelling rate, and tablet thickness change, we evaluated the distribution of each property. Histograms of each characteristic are shown in Fig. 1. TS and DT after the accelerated test are shown at a compression pressure of 160 MPa (Fig. 1). All of the characteristic showed a wide distribution, and it was confirmed that each characteristic varied significantly depending on the API. For example, focusing on TS after the accelerated test, ranitidine hydrochloride showed the lowest value of 0.23 MPa, indicating low tablet hardness. By contrast, neomycin sulfate showed the highest value of 13.98 MPa, indicating that it is a very hard tablet. In log*DT* after the accelerated test, acetylsalicylic acid showed the lowest value at 2.51, indicating that the tablets disintegrated quickly in water. By contrast, atenolol, probenecid, and gamma oryzanol had the lowest disintegration and did not disintegrate after 30 min.

Carbamazepine, probucol, and roxithromycin showed the lowest tablet thickness change of approximately 1.5%, with almost no change in tablet thickness before and after the accelerated test. By contrast, ranitidine hydrochloride showed the largest tablet thickness change, with tablets expanding by 21% after the accelerated test. The second-largest change in tablet thickness was observed with sodium salicylate at 9.9%, and the median tablet thickness change was 3.2%, indicating that ranitidine hydrochloride had by far the largest tablet thickness change. Ranitidine hydrochloride is known to be hygroscopic,[34] and this property may have resulted in a significant change in tablet thickness and lower TS after accelerated test.

A typical example of swelling behavior is shown in Fig. 2. The swelling behavior differed greatly among the APIs.

For example, D-naproxen and theophylline swelled rapidly, whereas L-valine swelled slowly and calcium pantothenate hardly swelled at all. Both D-naproxen and theophylline swelled quickly, but the maximum swelling force differed significantly, with D-naproxen having a greater maximum swelling force than theophylline. Among the APIs measured in this study, naproxen showed the highest maximum swelling force of 54 N, and calcium pantothenate showed the lowest swelling force of 1.2 N. Mebendazole showed the shortest swelling time and isoniazid the longest. Mebendazole swelled the fastest and isoniazid the slowest.

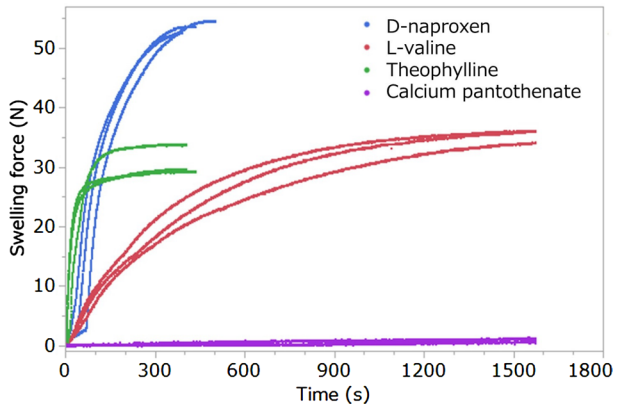


Fig. 2. Typical Example of Tablet Swelling Behavior

The results shown are individual values from three repeated measurements.

**Correlation between Properties** To evaluate how TS and log DT varied before and after the accelerated test, scatter plots were prepared for each tablet property before and after the accelerated test (Fig. 3). The coefficient of TS determination before and after the accelerated test showed 0.491–0.657. For the DT, the coefficient of determination was 0.696–0.728. The results were linear to some extent, but there were also many samples that deviated from the straight line (Fig. 3). These results confirm that although the properties were generally similar before and after the accelerated test, there was some degree of variation, and the properties changed before and after the accelerated test.

To confirm whether the newly added material properties can evaluate APIs from a different perspective, we evaluated the correlation with the property values obtained so far. The correlation coefficients between all properties are shown in Fig. 4. In most cases, the correlations between properties are low. Even in the case of the highest correlation, the correlation coefficient was −0.720, but in most cases, the correlation between properties was low. In addition, a principal component analysis was performed. The results of the analysis with principal components showed that up to the third principal component explained 55.3% of the data when the previously constructed database was used. By contrast, with the addition of the four new material properties, the amount of data that could be explained by the third principal component dropped

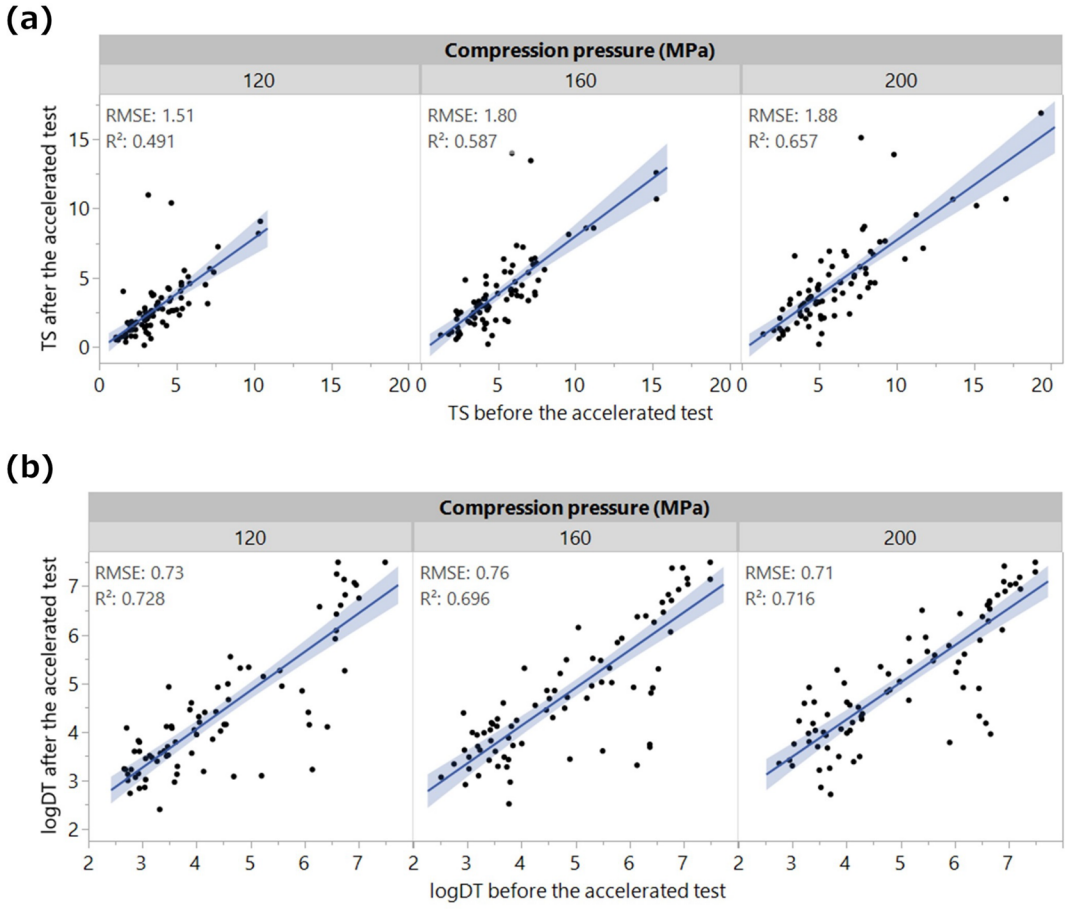


Fig. 3. Scatter Plot of (a) TS and (b) Log *DT* before and after the Accelerated Test, Respectively

Each point represents a specific API, and the figure shows the results of the tablets studied for 81 different APIs. All tablet formulations are identical. Regression lines and their 95% confidence intervals were calculated by single regression analysis.

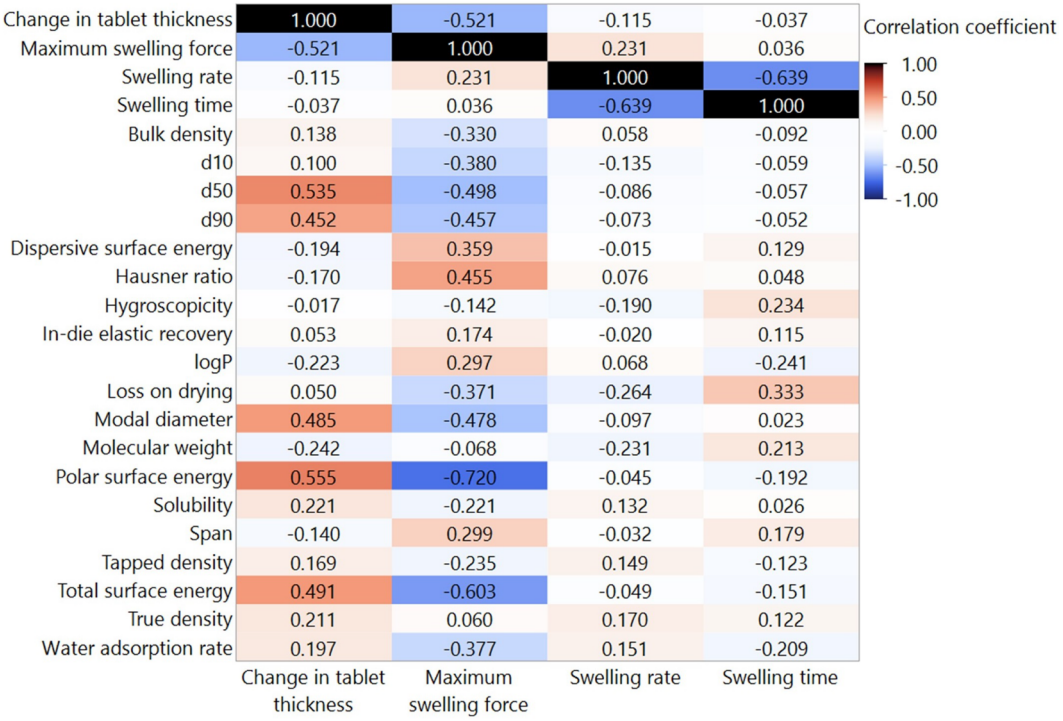


Fig. 4. Correlation Matrix between Material Properties

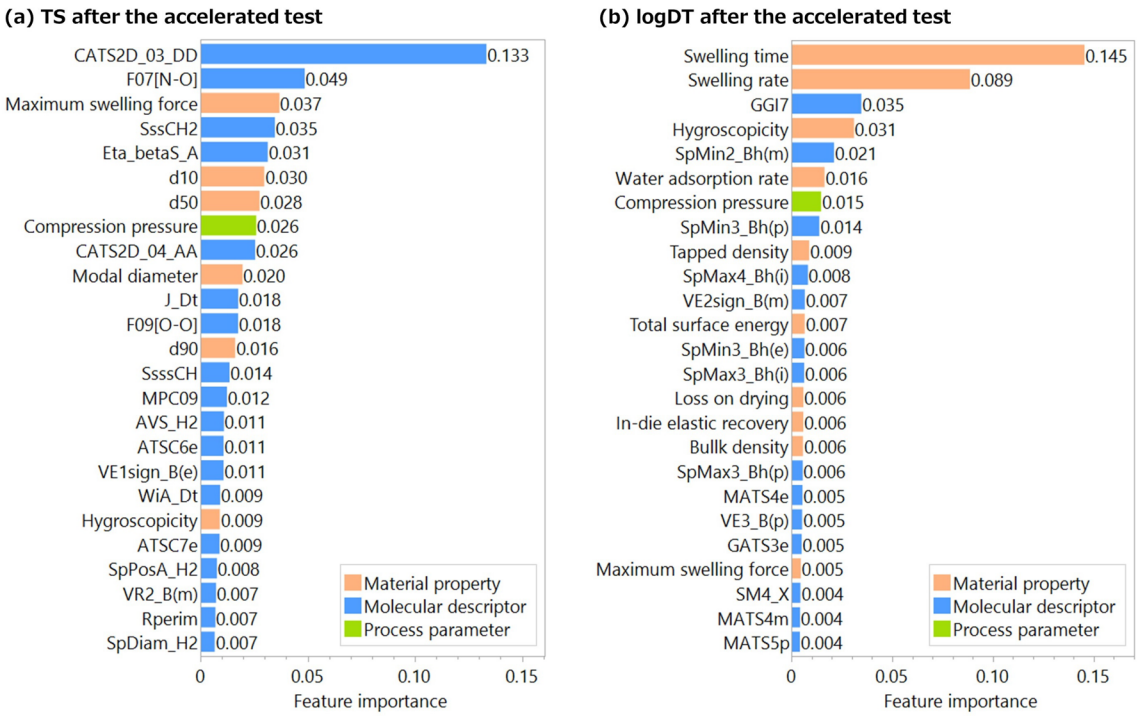


Fig. 5. Feature Importance for (a) TS and (b) Log*DT* after the Accelerated Test Calculated Using Random Forest

Of the 3,406 feature types, only 25 results with the highest feature importance are shown.

to 51.5%. This suggests that more dimensions are needed to explain the data and that the database could be expanded by adding new material properties. The features that contributed significantly to each principal component were as follows: first principal component: *d*10, mode diameter, maximum swelling force; second principal component: log*P*, total surface energy, polar surface energy; third principal component: molecular weight, loss on drying, swelling time.

**Feature Importance** Figure 5 shows the top 25 features in terms of feature importance for TS and log*DT* after the accelerated test calculated by RF. Of the 25 features for TS after the accelerated test, 18 were molecular descriptors, six

were material properties, and one was a process variable. For log*DT* after the accelerated test, 14 were molecular descriptors, 10 were material properties, and one was a process variable. Molecular descriptors had a significant effect on both tablet properties. Because after the accelerated test there were slightly more important molecular descriptors for TS than for log*DT*, it is likely that TS after the accelerated test was more closely related to the molecular descriptors. In particular, for TS after the accelerated test, the top two features were molecular descriptors, whereas for log*DT* after the accelerated test, the top two features were material properties.

As far as molecular descriptors are concerned, it is difficult to consider the mechanism of their effect on tablet properties. By contrast, we can infer to some extent the reason that material properties and compression pressure affect tablet properties. In TS after the accelerated test, features other than molecular descriptors that made it into the top 25 features were maximum swelling force, *d*10, *d*50, modal diameter, compression pressure, and hygroscopicity. The *d*10, *d*50, and modal diameter are physical properties expressing particle size. In general, the finer the particle size, the higher the specific surface area and the more contact points between particles, resulting in higher hardness. Besides its significant effect on TS before the accelerated test,[21] the particle size is also thought to have a significant effect on TS after the accelerated test because there is some correlation between TS before and after the accelerated test. In general, the higher the compression pressure, the less porosity there is in the tablet and the more contact points there are between the particles, resulting in higher hardness. The higher the maximum swelling force, the more the tablets swelled during the accelerated test and the more the voids increased, which is believed to break the bonds between the particles and make them more brittle. In addition, a higher hygroscopicity is believed to have affected the hardness of the tablets by encapsulating more water in the tablets, thereby promoting tablet swelling.

As for log*DT* after the accelerated test, the swelling time, swelling rate, hygroscopicity, water adsorption rate, compression pressure, tapped density, total surface energy, loss on drying, in-die elastic recovery, bulk density, and maximum swelling force were found to be important features. There are two mechanisms of tablet disintegration, swelling and wicking. Swelling is a phenomenon in which particles expand in all directions and push adjacent components apart, thereby breaking up the tablet matrix.[35] In general, the faster the swelling rate, the higher the swelling force, and the shorter the swelling time, the shorter the DT. For example, Colombo *et al.* prepared 21 different tablets with different API types, formulations, and tableting pressures, and evaluated DT and swelling behavior.[33] They found a negative correlation between DT and swelling rate. Wicking is a phenomenon in which liquid enters the microscopic pores in the tablet by capillary action and displaces air.[35] Therefore, it is reasonable that the swelling time, swelling rate, and water adsorption rate had a significant impact. The surface free energy is also strongly related to the wettability of the particle surface and has been reported to be related to DT in the oral cavity.[36] The higher the compression pressure, the less void space and the higher the tablet hardness, which may have affected the DT by making it more difficult for water to penetrate the tablet. Hygroscopicity was also important in the DT before the accelerated test, and the higher the hygroscopicity, the slower the DT. This is thought to be caused by extremely high hygroscopicity results in high particle-water bonding, creating a particle–water–particle bond that prevents particles from dispersing into the water when water infiltrates. As for the other material properties, the detailed mechanism is not known, but it is assumed that the accelerated test may be involved in the phenomenon, as the swelling and hydraulic conductivity may be increased or decreased by the accelerated test, thereby changing the DT.

To evaluate how feature importance changes before and after the accelerated test, a similar study was conducted on TS and log*DT* before the accelerated test (Supplementary Fig. S1). When we focus on the ordinal order of material properties and compaction pressure, both tablet properties are similar to our previous results,[21] but for log*DT*, the material properties related to swelling behavior, which were newly added in this study, had the strongest effect. The results in Supplementary Fig. S1 also show that molecular descriptors have a significant effect on TS and log*DT* even before the accelerated test. However, the order of the important features was found to change slightly. For TS, the rankings of particle size and compaction pressure increased, while the ranking of maximum swelling force and hygroscopicity decreased. For log*DT*, the swelling rate became by far the most influential factor, and the contribution of maximum swelling force also increased, whereas the ranking of hygroscopicity decreased. The change in feature importance before and after the accelerated test may be because of the tablets absorbing moisture during the accelerated test, which changed the internal structure of the tablets and the properties of the particles. Therefore, it is a plausible result that the importance of hygroscopicity has decreased.

**Evaluation of Prediction Accuracy of the Model** Figure 6 shows the prediction accuracy of the test set for each ML type, when the top 25 features including molecular descriptors, material properties, and compression pressure are used as explanatory variables. For both tablet properties after the accelerated test, BNN showed the best prediction accuracy. The model constructed using BNN showed coefficients of determination of 0.850 and 0.873 for TS and log*DT* after the accelerated test, respectively. RMSE showed 1.16 MPa and 0.503 for TS and log*DT* after the accelerated test, respectively, indicating that it can predict with high accuracy. Other ML models could also be constructed with a certain degree of high accuracy. With respect to regularized regression models such as elastic net, LASSO, and Ridge regression, in most cases, models with squared and interaction terms were more predictive than those with main effects only. In other words, the relationship between features and tablet properties was nonlinear and complex, with antagonistic and synergistic effects occurring. Our previous studies have shown that the interaction between material properties and tableting pressure, as well as the square term, affect tablet properties,[37] which is also consistent with the trend. With respect to the variability in prediction accuracy, the regularized regression model for TS showed more variability than the other ML methods. This indicates that the combination of training, validation, and test sets could vary prediction accuracy significantly. However, we were unable to determine what characteristics of the data set caused the significant decrease in prediction accuracy.

The prediction accuracy of the ML model that did not include molecular descriptors and consisted of 25 features

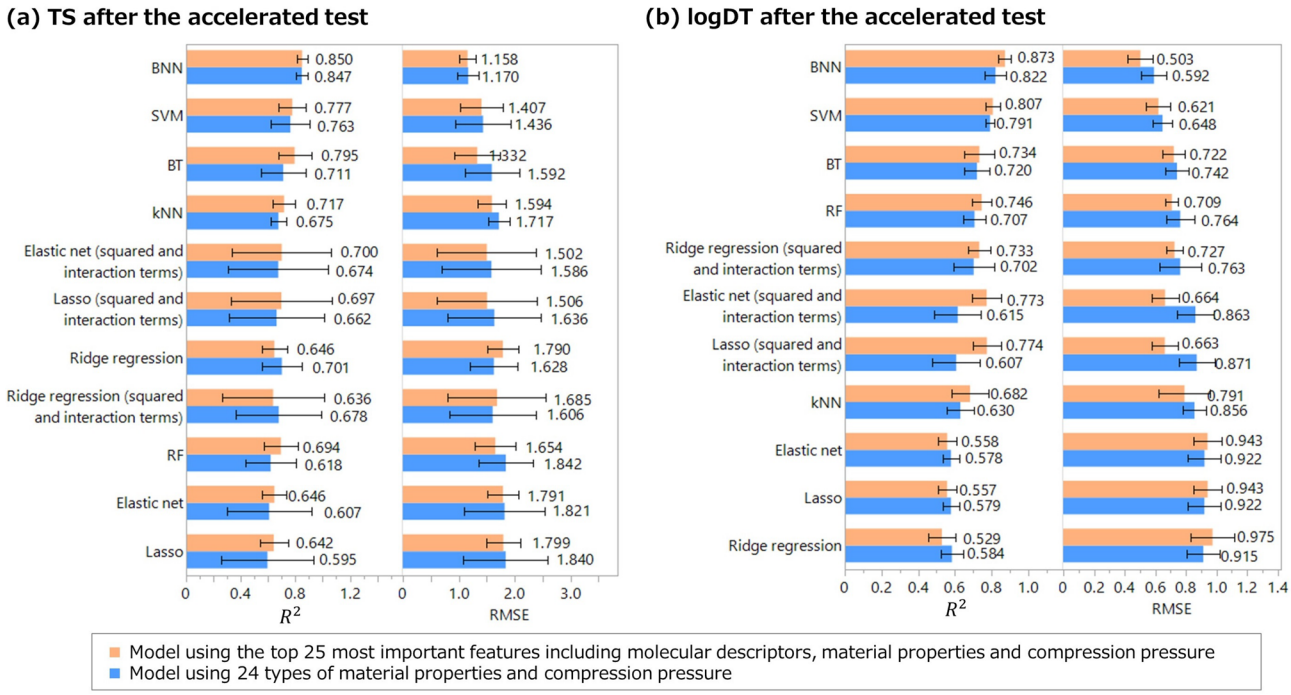


Fig. 6. Prediction Accuracy of (a) TS and (b) Log*DT* after the Accelerated Test in Each ML Method

Prediction accuracy was calculated using a test set consisting of unknown sample data. The prediction accuracy for each test set was calculated five times in different combinations. The figure shows the mean and standard deviation.

of 24 material properties and compression pressure was calculated, and the prediction accuracy was compared with a model that included molecular descriptors (Fig. 6). As with the models containing molecular descriptors, BNN showed the best prediction accuracy. The BNN model without molecular descriptors showed coefficients of determination of 0.847 and 0.822 for TS and log*DT* after the accelerated test, respectively. RMSE showed 1.17 MPa and 0.592 for TS and log*DT* after the accelerated test, respectively. The prediction accuracy for TS was almost equal to the model with molecular descriptors, and for log*DT*, the model with molecular descriptors was slightly more accurate than the model without. In other words, this result indicates that the material properties in this database could be substituted with molecular descriptors. Material properties are mainly based on experimentally measured physical properties, which are often very costly to acquire. Molecular descriptors, on the other hand, are computationally calculated values and can be easily obtained. Incorporating molecular descriptors into the model eliminates the need to measure several material properties and reduces the cost of constructing a tablet property prediction model.

As tablet hardness is generally affected by particle properties such as the van der Waals forces between particles and the DT value is affected not only by wettability, swelling, and wicking property but also by strain recovery and heat of interaction,[24] molecular descriptors may be related to these property values. For example, it has been reported that a surface property such as the water contact angle for APIs may be related to molecular descriptors.[38] Furthermore, we have shown that the true density of the API is strongly correlated with the molecular descriptor.[22] Therefore, with respect to other particle properties, they may also be closely related to the molecular structure. In particular, the addition of molecular descriptors for DT slightly improved the prediction accuracy, suggesting that there is a relationship between molecular descriptors and material properties that are not included in our database.

The prediction accuracy was also evaluated for TS and log*DT* before the accelerated test (Supplementary Fig. S2). The tablet properties before the accelerated test were also predicted with high accuracy. For TS before the accelerated test, SVM, BNN, and BT had the highest prediction accuracy, in that order. For logDT, BNN, SVM, and BT had the highest prediction accuracy, in that order. Similar to the results after the accelerated test, the model including molecular descriptors could predict tablet properties with high accuracy, indicating the possibility of reducing the number of material properties to be measured. However, in some cases, models consisting only of material properties and compression pressure had a slightly higher prediction accuracy. This may suggest that the molecular descriptors are more relevant to the properties of the tablets after the accelerated test.

**Relationships between the Number of Features and Prediction Accuracy** We evaluated how much we could reduce the number of features to be included in the prediction model by. Based on the results in Fig. 5, we gradually increased the number of features with high importance and evaluated the prediction accuracy. This time, we added three features each and evaluated the prediction accuracy of the test set. Figure 7 shows the relationships between the number of features and the prediction accuracy. For prediction accuracy calculations, only one test set was examined as an example. The prediction accuracy of the test set reached its head at approximately 15–18 features for both tablet properties. In other words, seven of the 25 features contribute little to the prediction. In the future, the measurement of less important features can be omitted when adding new APIs and updating the database by identifying the features necessary for prediction in this way,

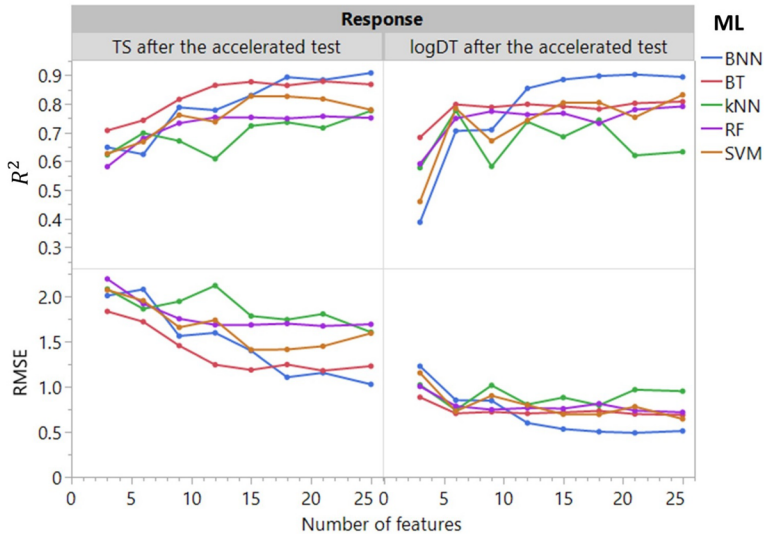


Fig. 7. Relationship between the Number of Features and the Prediction Accuracy of the Test Set

Only ML five types with particularly high prediction accuracy are shown.

thereby helping to improve the efficiency of constructing prediction models.

## Conclusion

We constructed a new material library that included 81 model APIs. It included 3381 types of molecular descriptors, 24 types of API material properties, and the TS and DT values obtained with three different compression pressures before and after the accelerated test. The RF showed that not only material properties and tableting pressure, but also molecular descriptors have a significant impact on TS and DT after the accelerated test. Compared with the model with 24 types of material properties and compression pressure, the model using the 25 most important features including molecular descriptors, material properties, and compression pressure showed comparable prediction accuracy with respect to TS after the accelerated test. With respect to DT after the accelerated test, the model with molecular descriptors had a higher prediction accuracy than the one without molecular descriptors. It was also shown that a model with approximately 15–18 important features is sufficient for building a highly accurate model, and that further increasing the number of features does not significantly improve the prediction accuracy. Molecular descriptors are obtained computationally, and thus have the potential of reducing the cost of measuring the material properties needed to build a predictive model of tablet properties after the accelerated test. This study demonstrated that a data-driven approach was effective in discovering complex relationships hidden in complex, large data sets and in predicting tablet properties after the accelerated test.

**Acknowledgments** This study was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Nos. 20K06986 and 22K15261. The authors thank Mr. Naohiro Masukawa, SAS Institute Japan Ltd., for his advice on ML and usage of the JMP Pro (SAS Institute Inc.).

**Conflict of Interest** The authors declare no conflict of interest. The Laboratory of Pharmaceutical Technology, University of Toyama is an endowed department, supported by an unrestricted grant from Nichi-Iko Pharmaceutical Co., Ltd. (Toyama, Japan).

**Supplementary Materials** This article contains supplementary materials.

## References

1) Yu L. X., *Pharm. Res.*, **25**, 781–791 (2008).
2) Hayashi Y., Shirotori K., Kosugi A., Kumada S., Leong K. H., Okada K., Onuki Y., *Pharmaceutics*, **12**, 601 (2020).
3) Yu L. X., Amidon G., Khan M. A., Hoag S. W., Polli J., Raju G. K., Woodcock J., *AAPS J.*, **16**, 771–783 (2014).
4) Wang Z., Sun Z., Yin H., Liu X., Wang J., Zhao H., Pang C. H., Wu T., Li S., Yin Z., Yu X., *Adv. Mater.*, **34**, 2104113 (2022).
5) Himanen L., Geurts A., Foster A. S., Rinke P., *Adv. Sci.*, **6**, 1900808 (2019).
6) Huang J., Kaul G., Cai C., Chatlapalli R., Hernandez-Abad P., Ghosh K., Nagi A., *Int. J. Pharm.*, **382**, 23–32 (2009).
7) Liu H., Galbraith S. C., Ricart B., Stanton C., Smith-Goettler B., Verdi L., O'Connor T., Lee S., Yoon S., *Int. J. Pharm.*, **525**, 249–263 (2017).
8) Tomba E., Facco P., Bezzo F., Barolo M., *Int. J. Pharm.*, **457**, 283–297 (2013).
9) Van Snick B., Dhondt J., Pandelaere K., Bertels J., Mertens R., Klingeleers D., Di Pretoro G., Remon J. P., Vervaet C., De Beer T., Vanhoorne V., *Int. J. Pharm.*, **549**, 415–435 (2018).
10) Wang Z., Cao J., Li W., Wang Y., Luo G., Qiao Y., Zhang Y., Xu B., *Sci. Rep.*, **11**, 1–13 (2021).
11) Lou H., Lian B., Hageman M. J., *J. Pharm. Sci.*, **110**, 3150–3165 (2021).
12) Yang Y., Ye Z., Su Y., Zhao Q., Li X., Ouyang D., *Acta Pharm. Sin. B*, **9**, 177–185 (2019).
13) Roggo Y., Jelsch M., Heger P., Ensslin S., Krumme M., *Eur. J. Pharm. Biopharm.*, **153**, 95–105 (2020).
14) Lou H., Chung J. I., Kiang Y. H., Xiao L. Y., Hageman M. J., *Int. J. Pharm.*, **555**, 368–379 (2019).
15) Onuki Y., Kawai S., Arai H., Maeda J., Takagaki K., Takayama K., *J. Pharm. Sci.*, **101**, 2372–2381 (2012).
16) Chaves M. K., Kelly R. C., Milne J. E., Burke S. E., Chaves M. K., Kelly R. C., Milne J. E., Burke S. E., *Pharm. Dev. Technol.*, **27**,

511–524 (2022).
17) Paul S., Baranwal Y., Tseng Y. C., *Int. J. Pharm.*, **599**, 120439 (2021).
18) Galata D. L., Könyves Z., Nagy B., Novák M., Mészáros L. A., Szabó E., Farkas A., Marosi G., Nagy Z. K., *Int. J. Pharm*., **597**, 120338 (2021).
19) Takayama K., Kawai S., Obata Y., Todo H., Sugibayashi K., *Chem. Pharm. Bull*., **65**, 967–972 (2017).
20) Hayashi Y., Oishi T., Shirotori K., Marumo Y., Kosugi A., Kumada S., Hirai D., Takayama K., Onuki Y., *Drug Dev. Ind. Pharm*., **44**, 1090–1098 (2018).
21) Hayashi Y., Nakano Y., Marumo Y., Kumada S., Okada K., Onuki Y., *Int. J. Pharm.*, **609**, 121158 (2021).
22) Hayashi Y., Marumo Y., Takahashi T., Nakano Y., Kosugi A., Kumada S., Hirai D., Takayama K., Onuki Y., *Int. J. Pharm*., **558**, 351–356 (2019).
23) Takagaki K., Arai H., Takayama K., *J. Pharm. Sci.*, **99**, 4201–4214 (2010).
24) Desai P. M., Liew C. V., Heng P. W. S., *J. Pharm. Sci*., **105**, 2545–2555 (2016).
25) Mendez K. M., Reinke S. N., Broadhurst D. I., *Metabolomics*, **15**, 1–15 (2019).
26) Díez-Sanmartín C., Sarasa Cabezuelo A., *J. Clin. Med*., **9**, 572 (2020).
27) Tohry A., Yazdani S., Hadavandi E., Mahmudzadeh E., Chelgani S. C., *Powder Technol.*, **381**, 280–284 (2021).
28) Ogutu J. O., Schulz-Streeck T., Piepho H.-P., *BMC Proc.*, **6**, S10 (2012).
29) Hastie T., Tibshirani R., Friedman J., "The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition," Springer-Verlag, New York, NY, 2009.
30) Le T., Epa V. C., Burden F. R., Winkler D. A., *Chem. Rev.*, **112**, 2889–2919 (2012).
31) Bergström C. A. S., Charman W. N., Porter C. J. H., *Adv. Drug Deliv. Rev.*, **101**, 6–21 (2016).
32) Pitt K. G., Newton J. M., Stanley P., *J. Mater. Sci*., **23**, 2723–2728 (1988).
33) Colombo P., Caramella C., Conte U., La Manna A., Guyot-Hermann A. M., Ringard J., *Drug Dev. Ind. Pharm.*, **7**, 135–153 (1981).
34) Khan S., Giradkar P., Yeole P., *PDA J. Pharm. Sci. Technol.*, **63**, 226–233 (2009).
35) Markl D., Zeitler J. A., *Pharm. Res.*, **34**, 890–917 (2017).
36) Fukami J., Ozawa A., Yoshihashi Y., Yonemochi E., Terada K., *Chem. Pharm. Bull.*, **53**, 1536–1539 (2005).
37) Oishi T., Hayashi Y., Noguchi M., Yano F., Kumada S., Takayama K., Okada K., Onuki Y., *Int. J. Pharm.*, **577**, 119083 (2020).
38) Suihko E., Forbes R. T., Korhonen O., Ketolainen J., Paronen P., Gynther J., Poso A., *J. Pharm. Sci*., **94**, 745–758 (2005).